

Data Warehouse Grundlagen

Seminarunterlage

Version: 2.15



Dieses Dokument wird durch die ORDIX AG veröffentlicht.

Copyright ORDIX AG. Alle Rechte vorbehalten.

Alle Produkt- und Dienstleistungs-Bezeichnungen sind Warenzeichen oder eingetragene Warenzeichen der jeweiligen Firmen und beziehen sich auf Eintragungen in den USA oder USA-Warenzeichen.

Weitere Logos und Produkt- oder Handelsnamen sind eingetragene Warenzeichen oder Warenzeichen der jeweiligen Unternehmen.

Kein Teil dieser Dokumentation darf ohne vorherige schriftliche Genehmigung der ORDIX AG weitergegeben oder benutzt werden.

Adressen der ORDIX AG

Die ORDIX AG besitzt folgende Geschäftsstellen

ORDIX AG
Karl-Schurz-Straße 19a
D-33100 Paderborn
Tel.: (+49) 0 52 51 / 10 63 - 0
Fax.: (+49) 01 80 / 1 67 34 90

ORDIX AG
An der alten Ziegelei 5
D-48157 Münster
Tel.: (+49) 02 51 / 9 24 35 – 00
Fax.: (+49) 01 80 / 1 67 34 90

ORDIX AG
Welser Straße 9
D-86368 Gersthofen
Tel.: (+49) 08 21 / 507 492 – 0
Fax.: (+49) 01 80 / 1 67 34 90

ORDIX AG
Kreuzberger Ring 13
D-65205 Wiesbaden
Tel.: (+49) 06 11 / 7 78 40 – 00
Fax.: (+49) 01 80 / 1 67 34 90

ORDIX AG
Wikingerstraße 18-20
D-51107 Köln
Tel.: (+49) 02 21 / 8 70 61 – 0
Fax.: (+49) 01 80 / 1 67 34 90

Internet: <http://www.ordix.de>

Email: training@ordix.de

Inhaltsverzeichnis

1	Einführung	8
1.1	Geschichtliches	9
1.2	Motivation	10
1.2.1	Hintergrund.....	11
1.3	Ist-Situation	12
1.4	Modernes Informationssystem	13
1.5	Was ist ein Data Warehouse?	14
1.5.1	Multiple Datenquellen.....	15
1.5.2	Unternehmensspezifisch skalierbar	16
1.5.3	Universelle Abfragen möglich	17
1.5.4	Hochleistungsplattform für Reporting.....	18
1.5.5	Analysen (Vergangenheit/Zukunft).....	19
1.6	Ziele.....	20
1.7	Gründe für ein separates DWH.....	21
1.8	Abgrenzung zu OLTP.....	23
1.8.1	Abgrenzung zu OLTP-Anfragen.....	24
1.8.2	Abgrenzung zu OLTP-Daten.....	25
1.8.3	Abgrenzung zu OLTP-Anwender	26
1.8.4	Abgrenzung zu OLTP - Zusammenfassung.....	27
1.9	Definition	28
1.9.1	Definition nach Inmon.....	29
1.9.1.1	Themenorientierung	30
1.9.1.2	Integrierte Datenbasis	31
1.9.1.3	Persistente Datenbasis	32
1.9.1.4	Chronologisierte Daten.....	33
1.10	Anwendungsgebiete.....	34
2	Multidimensionales Datenmodell	36
2.1	Überblick	37
2.2	Normalisierung	38
2.2.1	Überblick	38
2.2.2	Normalisierung von Daten.....	39
2.2.2.1	Redundanzen	40
2.2.2.2	0. Normalform.....	41
2.2.2.3	1. Normalform.....	42
2.2.2.4	2. Normalform.....	44
2.2.2.5	3. Normalform.....	47
2.2.2.6	Zusammenfassung Normalformen	49
2.3	Kennzahlen	50
2.3.1	Additive Kennzahlen.....	51
2.3.2	Semi-Additive Kennzahlen	52
2.3.3	Nicht-Additive Kennzahlen	53
2.4	Dimensionen	54
2.4.1	Eigenschaften von Dimensionen.....	55
2.4.1.1	Einfache Hierarchie	56
2.4.1.2	Parallele Hierarchie	57
2.4.2	Aufbau Dimensionstabelle	58
2.4.2.1	Beispiel 1	59
2.4.2.2	Beispiel 2	60
2.4.3	Junk Dimension.....	61
2.4.3.1	Junk Dimension – Beispiel	62
2.5	Fakten- und Dimensionstabellen	63
2.5.1	Aufbau Faktentabelle	64
2.5.2	Varianten von Fakten	65
2.5.2.1	Ereignis Fakt.....	65
2.5.2.2	Snapshot Fakt	66
2.5.3	Beispiel Faktentabelle	67
2.6	DWH Datenmodelle	68

2.6.1	Starschema	69
2.6.1.1	Vor- und Nachteile	70
2.6.1.2	Abfragen im Starschema	71
2.6.2	Snow Flake Schema	72
2.6.2.1	Schematische Darstellung	73
2.6.2.2	Vorteile	74
2.6.2.3	Nachteile	75
2.6.2.4	Abfragen im Snow Flake Schema	76
2.6.2.5	Mischformen	77
2.6.3	Galaxy Schema - Überblick	78
2.7	Data Vault – Grundlagen	79
2.7.1	Data Vault – Elemente	80
2.7.2	Data Vault – Beispiel	81
2.7.3	Data Vault – Vorteile	82
2.7.4	Data Vault – Nachteile	83
2.8	Slowly Changing Dimensions	84
2.8.1	Slowly Changing Dimension Anwendungsbereiche	85
2.8.2	Slowly Changing Dimensions - Fachlicher Schlüssel	86
2.8.3	Typ 1 – keine Historisierung	87
2.8.4	Typ 2 – Historisierung	88
2.8.4.1	Typ 2 – Ablaufprinzip	89
2.8.4.2	Typ 2 – Beispiel	90
2.8.5	Typ 3 – Teilweise Historisierung	92
2.9	Designprozess	93
2.9.1	Überblick	93
2.9.2	Beispiel	94
3	Grundlagen der Architektur	96
3.1	Überblick	97
3.1.1	Architekturschichten	97
3.1.2	Schematischer Aufbau	98
3.2	ETL-Tools	99
3.3	Staging Area	100
3.4	Landing Area	101
3.5	Cleansing Area	102
3.6	Metadaten	103
3.7	Core Data Warehouse	104
3.8	Data Marts	105
3.8.1	Überblick	105
3.8.2	Extraktarten	106
3.8.3	Vorteile	107
3.8.4	Data Mart Arten	108
3.8.4.1	Abhängige Data Marts	108
3.8.4.2	Unabhängige Data Marts	109
3.8.4.3	Virtuelle Data Marts	110
4	Entwurf eines Data Warehouse Systems	111
4.1	Vorgehensmodell	112
4.2	Modellierungsschritte	114
4.3	Analyse des Informationssystems / Anforderungsanalyse	115
4.3.1	Informationsbedarfsanalyse	116
4.3.2	Analysemodell / Prozessmodell	119
4.3.3	Objektmodell	121
4.4	Konzeptioneller Entwurf	122
4.4.1	Beispiel MERM Diagramm	123
4.5	Logischer Entwurf	124
4.6	Technische Implementierung	125
4.7	Test	126
4.8	Softwareauswahl	127
4.8.1	Auswahlkriterien	128

4.8.2	Marktrecherche	129
4.8.3	Bewertung der Auswahl	130
4.8.4	Kosten der Software	131
5	Speicherstrukturen	132
5.1	Überblick	133
5.2	ROLAP	134
5.3	MOLAP	135
5.4	HOLAP	138
5.5	DOLAP	139
5.6	Multidimensional Expressions (MDX)	140
6	Befüllung	142
6.1	Überblick	143
6.2	ETL-Tool	144
6.3	Monitoring Quellsystem	145
6.3.1	Überblick	145
6.3.2	Triggerbasiert	146
6.3.3	Replikationsbasiert	147
6.3.4	Zeitstempelbasierte Monitoringstrategie	148
6.3.5	Log-basierte Monitoringstrategie	149
6.3.6	Snapshot-basierte Monitoringstrategie	150
6.4	Extraktionsstrategien	151
6.5	Staging Area	153
6.5.1	Überblick	153
6.5.2	Ausprägungen	155
6.6	Cleansing Area	156
6.7	Transformation	157
6.7.1	Ursache von fehlerhaften Daten	157
6.7.2	Überblick	158
6.7.3	Filterung	159
6.7.3.1	Überblick	159
6.7.3.2	Klassen	160
6.7.3.3	Beispiele	161
6.7.4	Harmonisierung	162
6.7.5	Aggregation	165
6.7.6	Anreicherung	166
6.8	Laden	167
6.9	Data Mart	168
6.10	Change Data Capture	169
6.10.1	Überblick	169
6.10.2	Vorteile	170
6.11	Deploymentprozess	171
6.11.1	Deploymentprozess – Aufbau der Testdatenbank	171
6.11.2	Deploymentprozess – ETL Prozesse Entwickeln	172
6.11.3	Deploymentprozess – Testen	173
6.11.4	Deploymentprozess – ETL Prozesse entwickeln	174
7	Multidimensionale Operatoren	175
7.1	OLAP Operatoren	176
7.1.1	Standard Operatoren	177
7.1.2	Bewegen im Multidimensionalen Datenmodell	178
7.1.3	Pivotierung/Rotation	179
7.1.4	Roll-up/Drill-down	180
7.1.4.1	Beispiel	181
7.1.5	Drill Across	182
7.1.6	Drill Through	183
7.1.7	Slice/Dice	184
7.1.8	Slice	185
7.1.9	Dice	186

7.1.10	Split / Merge	187
7.1.10.1	Beispiel	188
8	Reporting.....	189
8.1	Frontend Tools	190
8.2	Verteilung der Anwender.....	191
8.3	Dashboard.....	192
8.4	Statische Reports.....	193
8.5	Dynamische Berichte	194
8.6	Ad-Hoc Berichte	195
8.7	Data Mining	196
8.8	Auswahl Reporting Tool.....	197
9	Datenbankoptimierung	199
9.1	Überblick	200
9.2	Laden von Daten.....	201
9.2.1	Überblick	201
9.2.2	Einzelatzverarbeitung	202
9.2.3	Lade Tool	203
9.2.4	Beispiel Oracle	204
9.2.5	Beispiel DB2.....	205
9.2.6	Beispiel Informix	206
9.3	Partitionierung	207
9.3.1	Überblick	207
9.3.2	Range Partitionierung.....	208
9.3.3	List Partitionierung.....	212
9.3.4	Hash Partitionierung.....	215
9.4	Komprimierung.....	219
9.5	Datenbank Caches.....	222
9.6	Blockgröße – Seitengröße	223
9.7	Reservierter Freiplatz bei der Tabellenerstellung	224
9.8	Referenzielle Integrität	225
9.9	Materialisierte Sichten und Tabellen.....	226
9.10	Merge Anweisung	229
9.10.1	Beispiel.....	230
9.11	Parallelisierung.....	231
9.12	Spaltenorientierte Speicherung und In-Memory-Funktionalität.....	232
9.12.1	Überblick	232
9.12.2	Beispiel Oracle	233
9.12.3	Beispiel DB2.....	237
9.13	Hardwareoptimierungen.....	238
10	Big Data	239
10.1	Größe ist relativ	240
10.2	Potential	241
10.3	Wirkprinzipien.....	242
10.4	Die 3 V's.....	243
10.4.1	Velocity (Geschwindigkeit)	243
10.4.2	Volume (Datenmengen)	244
10.4.3	Variety (Vielfalt).....	245
10.5	Big Data ist nicht nur ein Tool	246
10.6	Big Data in Aktion.....	247
11	Übungen / Lösungen.....	249
11.1	Übungen.....	250
11.1.1	Normalisierung von Daten.....	250
11.1.2	Multidimensionales Datenmodell	253
11.1.3	Snowflakeschema	255
11.1.4	Befüllung	256
11.2	Lösungen.....	258

11.2.1	Normalisierung von Daten.....	258
11.2.2	Multidimensionales Datenmodell	262
11.2.3	Snowflakeschema	264
11.2.4	Befüllung	265
Anhang-	ETL Tools.....	267
Anhang-	Reporting Tools	277